What makes Sparkify users churn?

Introduction

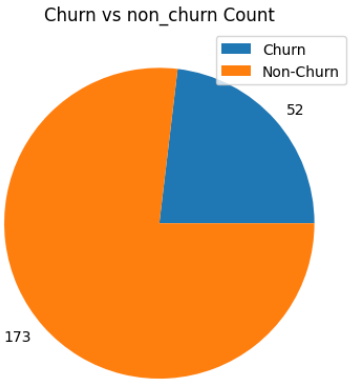
Our day-to-day life is full of stresses. One thing I like to do to disconnect from my stresses is to listen to music. Sparkify is now one of the most popular music streaming applications.

Last month, Sparkify launched a competition for data scientists to build a machine learning model that can accurately predict their users churn patterns so that they could approach those users with attractive promotions and retention offers. They published a dataset that includes interesting details about users' activities on the app during October and November 2018. So, why do users churn?

Analyzing the data

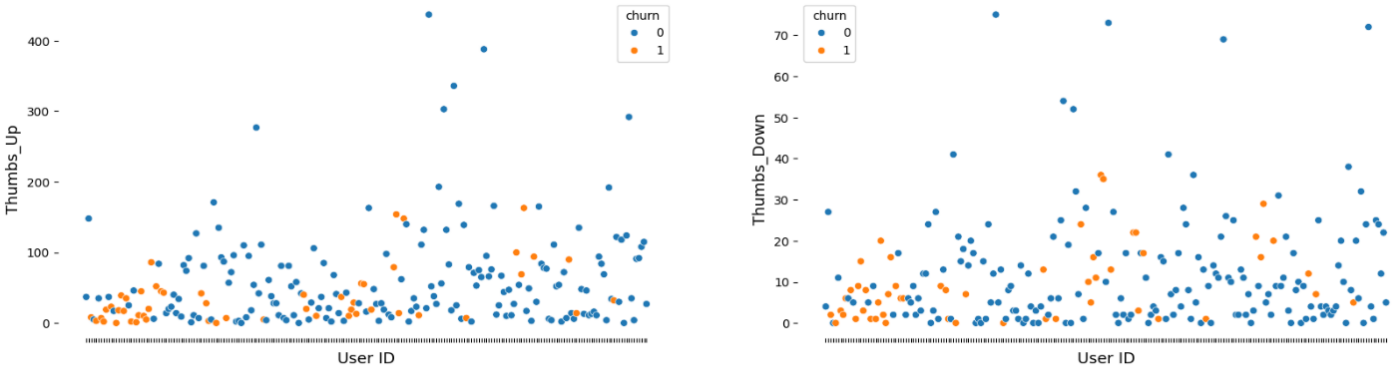
I started by analyzing the data to understand users' behavior. Here is what I found out:

- The dataset includes activities for 225 unique users, below is their churn distribution.



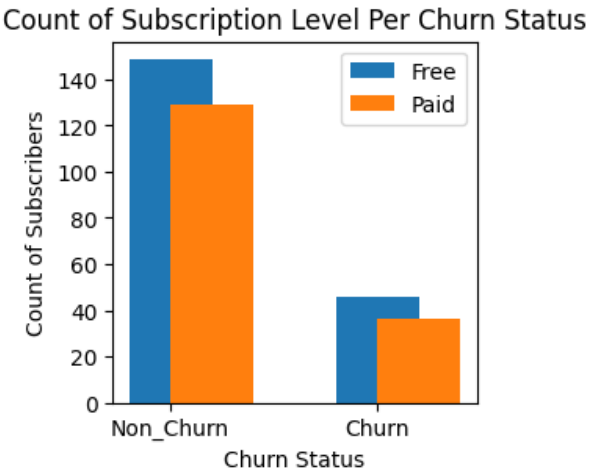
As we can see, the dataset is imbalanced.

- There is no column for churn status, I had to create it and add values based on user activity when they press the “Cancellation Confirmation” page. This page is one of many other pages that exist in the “page” column. Other values in this column include “Thumbs up”, “Thumbs down” which customers use to like or dislike a song, also we have “Add to playlist”, “Add friend”, and we have “NextSong” which indicates that a user is currently listening to a song, and many more values. The below chart shows counts for both churn and non-churn users.

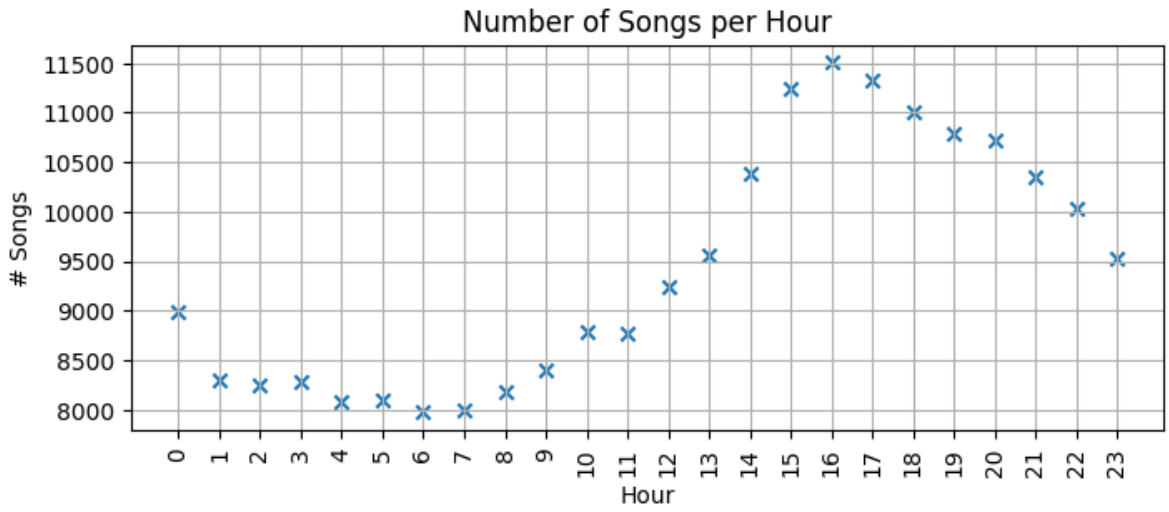


We notice that churn users (Orange dots) have made “Thumbs_down” more often than “thumbs_ up”.

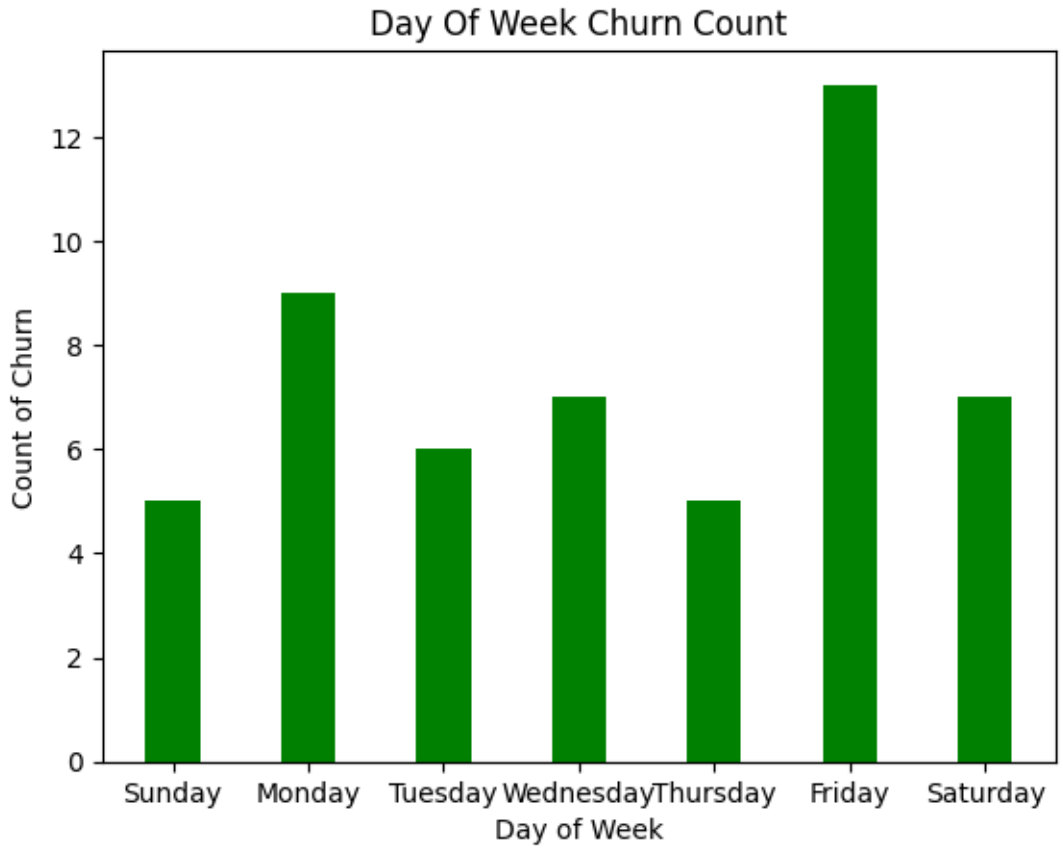
- The application has 2 subscription levels, free and paid. Let’s see the number of churn cases in both levels.



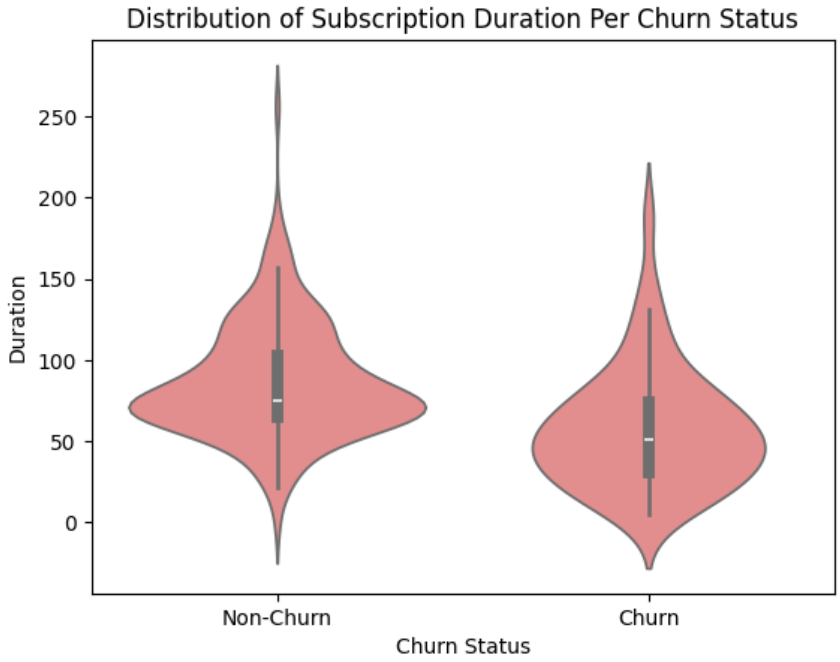
- I calculated the number of songs played per hour, and the chart shows that the peak time is from 3:00 PM to 6:00 PM



- One question that I was eager to answer is, which day of the week has the highest number of churn cases, the answer was found to be Friday as shown in the below chart.

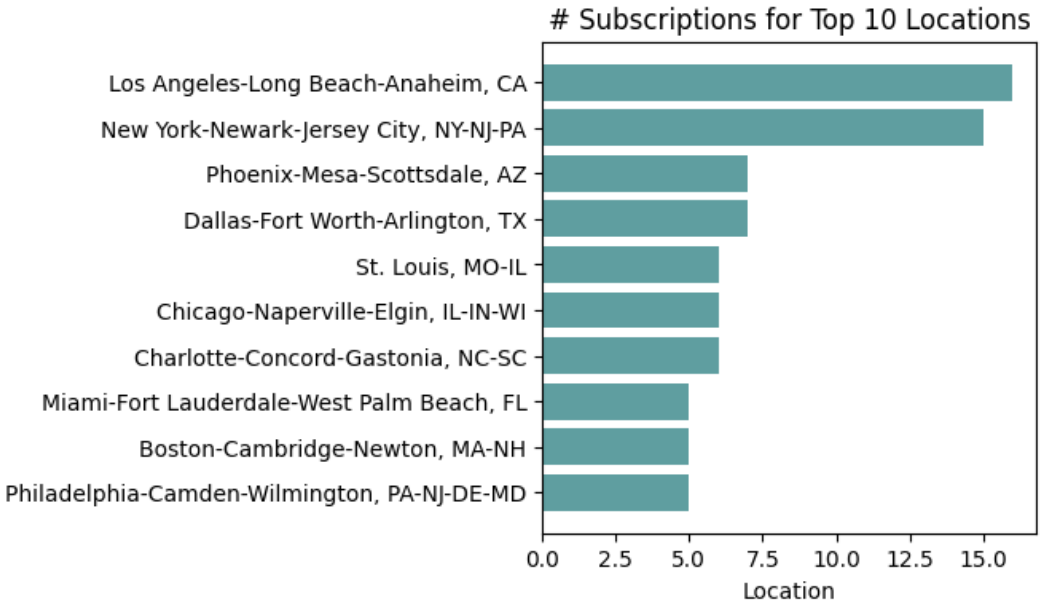


- I wanted to explore the distribution of the duration, which is how many days of subscription each user has, we can find the answer in the below violin plot.

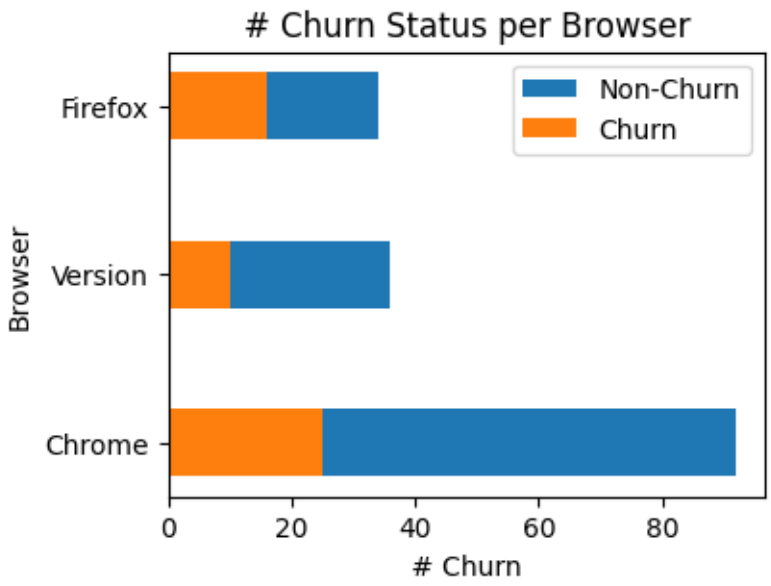


The plot shows that churn users tend to have shorter subscription duration.

- One other interesting information is about users' locations, the two locations with the highest number of subscribers appeared to be Los Angeles & New York.



- I also looked at the browsers used by customers, and here are the 3 types of browsers found.



Modelling the data

After preparing and analyzing the data, I built a machine learning model to predict churn cases. I tried 3 different classifiers: Gradient Boosting, Random Forest and Decision Tree. The best performance resulted from using **Random Forest.**

Since we have an imbalanced dataset, I used f1-score to evaluate our model performance.

I tried using grid-search cross-validation with different hyper-parameters values, but the best model performed slightly worse that the baseline model, so I used the baseline model with its default hyper-parameters, it returned an f1-score value 76.41% on our test set.

I also extracted feature importances and found that the two most important features to predict churn are: **# subscription duration days**, and **average song length**.

Conclusion

At this stage, we have a trained random forest model that can be used to predict churn cases.

We can use the information about the most important features to detect users with high probability of churning and try to approach them with retention offers.

Resources

The column description was obtained from a medium post
https://medium.com/@ferenc.hechler/sparkify-customer-satisfaction-prediction-f3b0941e8710

You can view my code here:
https://github.com/nesmasilman/Udacity_Data_Scientist_Nano_Degree_Sparkify_Capstone_Project